

## De la classification à la connaissance des maladies : une réflexion à partir de la classification internationale des maladies (CIM)

G. Botti<sup>1</sup>, O. Bodenreider<sup>3</sup>, A. Burgun<sup>2</sup>, P. Le Beux<sup>2</sup>, M. Fieschi<sup>1</sup>

<sup>1</sup>SIM, hôpital Timone-Adultes, 264, rue Saint-Pierre, 13385 Marseille cedex 5 ; <sup>2</sup>DIM, hôpital Pontchaillou, 35033 Rennes cedex, France ; <sup>3</sup>National Library of Medicine, Bethesda, Maryland, États-Unis

(Reçu le 1<sup>er</sup> mars 2000 ; révisé le 1<sup>er</sup> juin 2000 ; accepté le 1<sup>er</sup> septembre 2000)

### Résumé

Nous avons besoin de systèmes de classification tels que la CIM-10, constitués de classes prédéfinies, indépendantes et exclusives auxquelles affecter des situations cliniques pour élaborer des statistiques descriptives. Le Programme de médicalisation des systèmes d'information est à l'origine, en France, de la constitution de vastes bases de données médicales codées de cette façon, mais dont l'objectif est d'abord financier. Malgré cela, est-il possible de retrouver une information de type épidémiologique sur les maladies dans ces bases ? C'est loin d'être évident. Des éléments de discussion sont proposés à ce sujet. © 2000 Éditions scientifiques et médicales Elsevier SAS

**classifications / maladies / représentation de la connaissance**

### Summary – From statistic-oriented classification to disease knowledge: reflections on ICD-10.

*There is a need for statistical classification (like CIM-10) to provide purpose-dependant predefined classes where cases of interest have to be uniquely assigned. PMSI in France leads to a large medical database encoded with such a classification, whose aim is also financial. In spite of this, is it possible to retrieve disease information for epidemiologic descriptions? Not very easy in fact! This paper discusses the problem. © 2000 Éditions scientifiques et médicales Elsevier SAS*

**classification systems / diseases / representation of knowledge**

## INTRODUCTION

La classification est un mode de représentation de la connaissance largement utilisé dans le domaine médical. Avantages et inconvénients ont été bien étudiés [1-3] : nous avons besoin de vocabulaire proposant des classes prédéfinies adaptées aux objectifs des recueils d'information, où chaque situation à décrire trouve sa place et uniquement celle-là. Mais nous savons également que classer c'est toujours perdre de l'information – on choisit une façon de ranger l'information, donc on laisse de côté des

alternatives. Par ailleurs, seul sera décrit le vocabulaire utile dans le contexte où la classification est utilisée : c'est le principe de parcimonie. De ce fait, la « couverture » du domaine est généralement partielle. La Classification internationale des maladies (CIM) [4] est certainement la classification de maladies la plus utilisée dans le monde. En France, elle sert au codage des diagnostics dans le cadre du Programme de médicalisation des systèmes d'information (PMSI) [5] qui concerne l'ensemble des séjours hospitaliers. L'allocation budgétaire est le principal objectif de la mise en place de cette

volumineuse base d'information, mais bien d'autres usages pourraient en être faits et tous convergent vers le besoin d'améliorer notre connaissance de la population qui passe dans nos hôpitaux et des soins qu'elle y reçoit. Or la CIM ne contient qu'un aspect parcellaire de la connaissance des maladies, très influencé par son propre objectif de statistiques de morbidité. Couplée aux règles de codage liées à l'objectif budgétaire, elle n'autorise pas aujourd'hui d'analyse satisfaisante concernant les pathologies [6, 7].

Dans une première partie, une analyse détaillera la CIM-10 sous l'angle de la représentation de la connaissance des maladies. Dans une deuxième partie, c'est au sein des résumés par unité médicale (RUM) que ces dernières seront recherchées. Puis, mettant à profit les résultats issus de ces analyses, à propos d'un exemple sur le cancer, nous tenterons de conclure sur les possibilités que nous avons de parler de maladie dans notre système d'information médicalisé aujourd'hui.

## LES MALADIES DANS LA CIM

Classifier, c'est répartir des objets dans des classes selon un ou des critères de classification. Ces derniers sont fortement influencés par le système d'information dans lequel la classification est mise en œuvre. L'usage de la CIM en vue de la production de statistiques descriptives des états morbides est affirmé dans son titre lui-même « Classification statistique internationale » et, plus loin, dans l'introduction : « Une classification statistique des maladies doit couvrir tout l'éventail des états morbides avec un nombre raisonnable de catégories ». La fréquence des maladies recensées pèse indirectement sur la description, son usage international aussi : hypertrophie de certains aspects infectieux relativement à la pauvreté de la neurologie par exemple. Par ailleurs, tout « cas » doit pouvoir être classé, ce qui justifie les regroupements génériques de types « autres problèmes non classés ailleurs ». Une certaine ambiguïté transparait dès l'énoncé du champ d'application : « la CIM est utilisée pour transposer les diagnostics de maladies ou d'autres problèmes de santé, en codes alphanumériques, ce qui facilite le stockage, la recherche et l'analyse des données ». Diagnostic ? Maladie ? Quand on sait que dans le PMSI on ne parle également que de diagnostic, on comprend qu'il soit difficile de retrouver les maladies dans la base par la suite.

### Vocabulaire et types sémantiques

Concernant la description des maladies, on trouve dans la CIM les informations suivantes, plus ou moins explicites selon les cas.

#### *Pathologies proprement dites*

*Pathologies dites « d'organe »* pour certains chapitres tels que le chapitre III « Maladies du sang et des organes

hématopoïétiques et certains troubles du système immunitaire », le chapitre IV « Maladies endocriniennes, nutritionnelles et métaboliques », etc. Dans ce cas, l'organe ou le système concerné est toujours connu, même si cela peut être à un niveau assez général, par exemple [G99.8 Autre affection du système nerveux]. Il existe des exceptions : le chapitre IV fait bien référence à un organe (glandes endocrines) mais pas uniquement (nutrition et du métabolisme).

*Pathologies autres que d'organe* pour le chapitre I « Certaines maladies infectieuses et parasitaires », le chapitre II « Tumeurs », le chapitre XVII « Malformations congénitales et anomalies chromosomiques » et la plus grande partie du chapitre XIX « Lésions traumatiques, empoisonnement et certaines autres conséquences de causes externes ». Dans ce cas, c'est sur le processus pathologique au sens large que porte l'information. Au niveau du code à quatre caractères, l'organe est généralement explicite, mais pas toujours, par exemple [Q899 Malformation congénitale, sans précision].

*Cas particulier des nouveau-nés* (chapitre XVI) et de la grossesse (chapitre XV) : chapitres conçus pour être utilisés de façon autonome et qui contiennent des actes (groupe 080-084).

#### *Signes ou symptômes*

À peu près toujours relatifs à des organes ou des appareils et concentrés dans le chapitre XVIII « Symptômes, signes et résultats anormaux d'examen cliniques et de laboratoire », non classés ailleurs. L'information est alors explicite par la structure, par exemple [R51 Céphalées]. Cependant, ils peuvent être attachés au chapitre dont ils relèvent dès lors qu'ils sont très spécifiques et l'information est alors implicite, par exemple [L299 Prurit, sans précision] au sein du chapitre XII.

#### *Maladies compliquant une maladie ou son traitement*

La complication peut être explicitée par la structure : par exemple le code [T845 Infection et réaction inflammatoire due à une prothèse articulaire interne] est rattaché à la catégorie à trois caractères [T84 Complication de prothèses, implants et greffes orthopédiques internes]. Elle peut être explicite par le libellé : [E10.2 Diabète insulino-dépendant avec complications rénales]. Elle peut être implicite : pour le code [I12.0 Néphropathie hypertensive, avec insuffisance rénale], seul « avec » peut évoquer, mais évoquer seulement, la complication ; dans [D70 Agranulocytose] rien dans le libellé n'évoque la possibilité que cette maladie soit une complication. En revanche, dans les inclusions qui lui sont attachées, on trouve « neutropénie médicamenteuse » qui peut y faire penser.

#### *Notion de séquelles de maladies*

Regroupées au sein de sous-chapitres tels que « Séquelles de maladies infectieuses et parasitaires » (codes B90-

B94) ou « Séquelles de lésions traumatiques, d'empoisonnement et d'autres conséquences de causes externes » (codes T90-T98). L'information y est explicite par la structure et le libellé.

### Notion de thérapeutique iatrogène

Se trouve à plusieurs endroits. Les codes Y40-Y84 traduisent la notion de complication iatrogène médicamenteuse (exemple : [Y44.2 Anticoagulants]... ayant provoqué des effets indésirables...) ou chirurgicale (exemple : [Y71 Appareils cardiovasculaires, associés à des accidents]).

Il faut bien noter que les codes de séquelles et de thérapeutiques iatrogènes ne correspondent pas à des maladies mais à la notion de complication ou de séquelle en tant que telle. Ils typent des situations et sont donc potentiellement très riches pour structurer le problème médical, d'autant plus que dans les bases de résumés, les codes traduisant ces deux informations (la cause et la conséquence) coexistent le plus souvent. Ainsi, par exemple, [D70 Agranulocytose] est la conséquence et [Y43.1 Antimétabolites antitumoraux] la cause. Pour le reste, la complication en tant que maladie, complication d'une maladie ou résultat de son traitement, possède généralement un code dans la CIM mais nous venons de voir que le lien entre les deux est réellement multiforme. C'est cela qui rend la CIM si difficile à utiliser comme base de connaissance. Un dernier exemple : [K913 Occlusion intestinale post-opératoire, non classée ailleurs] appartient à la catégorie à trois caractères [K91 Atteintes de l'appareil digestif, après un acte à visée diagnostique ou thérapeutique, non classés ailleurs] signifiant des complications iatrogènes que le terme « post- » peut impliquer. Seule les règles d'utilisation de la CIM nous permettent de savoir qu'il faut éviter dans cette situation clinique de coder [K56.6 Occlusion du colon ou de l'intestin] puisque [K913 Occlusion intestinale post-opératoire, non classée ailleurs] existe. Dans ce cas, il existe deux codes différents pour le même concept d'occlusion intestinale, selon qu'il est utilisé dans un contexte (maladie) ou un autre (complication). Dans d'autres cas, par exemple [D70 Agranulocytose], on utilisera le même code dans les deux cas.

En résumé, il se dégage de ce bref survol qu'à une grande partie des codes de la CIM, on peut affecter un des types sémantiques suivants :

- « signe-ou-symptôme, maladie », simple si l'entité se décline au minimum avec le doublet [processus pathologique] [localisation] ou complexe lorsque plusieurs doublets [processus pathologique] [localisation] sont nécessaires ;
- « complication, notion-de-séquelle, notion-de-complication », liée à un traitement médical ou à un acte.

Les codes de la CIM qui ne peuvent bénéficier d'un des types sémantiques ci-dessus ne font pas partie de la description des maladies. Par exemple, la notion de dépistage d'une maladie, qui correspond plus à une activité qu'à une

**Tableau I. Hiérarchie de la Classification internationale des maladies.**

| Modèle                             | Exemple                         |
|------------------------------------|---------------------------------|
| Chapitre                           | Maladies de l'appareil digestif |
| Groupe                             | Maladies du péritoine           |
| Catégorie à trois caractères       | Péritonite                      |
| Sous-catégorie à quatre caractères | Péritonite aiguë                |

**Tableau II. Organisation dans les hiérarchies.**

| Modèle                      | Exemple                                |
|-----------------------------|--|
| Niveau <i>n</i>             | Maladies de l'appareil génito-urinaire |
| Niveau <i>n</i> + 1 précis  | Insuffisance rénale                    |
| Niveau <i>n</i> + 2 précis  | Insuffisance rénale aiguë              |
| Niveau <i>n</i> + 2 général | Insuffisance rénale, sans précision    |
| Niveau <i>n</i> + 1 précis  | Lithiases urinaires                    |
| Niveau <i>n</i> + 1 général | Autres maladies de l'appareil urinaire |

maladie, ainsi que l'ensemble des motifs de recours aux soins qui ne sont pas des maladies, à plus forte raison quand ils sont des actes.

On trouve parfois, dispersées au fil des chapitres, des informations correspondant à des modalités évolutives (appendicite aiguë, pancréatite chronique) dans les libellés. Cela reste très modeste : la notion de récurrence ou de rechute, par exemple, n'est pas connue de la CIM. Mais on peut parfois l'apprécier indirectement : coexistence des codes [I252 Antécédent d'infarctus du myocarde] et [I219 Infarctus aiguë du myocarde]. De même, la notion de gravité n'est jamais explicite, sous la forme de stades évolutifs individualisés. Ainsi, lorsqu'une affection peut être soit légère, soit complètement invalidante, selon son stade évolutif, le seul moyen de coder la gravité est le regroupement de codes pour tenter de décrire la situation : [G35 Sclérose en plaque] et [R26.2 Difficulté à la marche, non classée ailleurs] par exemple.

### Hiérarchie et niveaux de granularité des concepts

Le niveau le plus général de la CIM est le chapitre, suivi du groupe, et parfois du sous-groupe comme dans le chapitre des tumeurs, puis de la catégorie à trois caractères et de la sous-catégorie à quatre caractères. Ce sont les liens implicites « fait partie de » ou « est un » qui constituent principalement les hiérarchies (tableau I). Au sein de chacun des niveaux de description, l'organisation générale des concepts est comparable, allant du précis vers le général (tableau II), c'est-à-dire au sein d'un chapitre, pour les groupes qui le composent (tuberculose... autres maladies bactériennes... autres maladies infectieuses par

exemple), au sein d'un groupe, pour les catégories qui le composent (choléra... autres infections intestinales par exemple), au sein d'une catégorie, pour les sous-catégories qui la composent (méningite tuberculeuse... tuberculose du système nerveux SAI, par exemple).

Cette structuration a pour conséquence que les codes à quatre caractères correspondent :

– soit à des concepts précis : [A15.6 Pleurésie tuberculeuse], [I21.1 Infarctus du myocarde paroi antérieure], etc. ;

– soit, à l'opposé, à des concepts très génériques : [B88.2 Autres infestations par des arthropodes], [I98.1 Troubles cardiovasculaires au cours d'autres maladies infectieuses et parasitaires classées ailleurs], etc., selon que l'on se situe au début ou à la fin d'un niveau de description.

Les niveaux de granularité des concepts ne sont pas superposables aux niveaux des hiérarchies des codes. On retrouve cela à l'analyse des codes ---.9 : le code [I36.9 Atteinte non-rhumatismale de la valvule tricuspide, sans précision] est un code correspondant à un concept assez précis, alors que le code [I73.9 Maladie vasculaire périphérique, sans précision] est très générique. On le voit aussi en comparant les libellés contenant la chaîne de caractères « sans précision ». Le code [B19.0 Hépatite virale sans précision, avec coma] montre bien toutes ces difficultés de description : « hépatite virale » est plus précis « qu'hépatite » mais moins « qu'hépatite aiguë B » par exemple. Le manque de précision porte bien sur le virus responsable, car par ailleurs la notion d'hépatite virale dans le coma correspond à un tableau déjà bien précis... La maladie est décrite selon les cas par les facettes suivantes : pathologie (ici inflammation), localisation (ici foie), étiologie (ici virale), mode d'évolution (aigu ou chronique, ici non précisé), gravité clinique (ici coma). Le fait de représenter l'information sur un seul axe dans la CIM amène à combiner, selon les libellés, certains des aspects qui constituent la maladie. Ces aspects peuvent ensuite être décrits de façon plus ou moins précise. Par ailleurs, si les codes ---.0 à ---.8 viennent préciser l'intitulé de la catégorie (relation d'hyponymie, par exemple [M54.2 Cervicalgie] « est une » [M54 Dorsalgies]), les codes ---.9 rompent cette organisation puisqu'ils sont au même niveau et donc redondants avec elle dans une grande majorité des cas ([M54.9 Dorsalgie, sans précision] n'est pas plus précis que [M54 Dorsalgies]). Il arrive que ce schéma soit mis en défaut : par exemple [I12.9 Néphropathie hypertensive, sans insuffisance rénale] et [I12 Néphropathie hypertensive]). On retiendra que les niveaux dans les hiérarchies de codes ne peuvent être utilisés pour déduire des degrés de précision des concepts.

### Hiérarchie et proximité sémantique

C'est le propre d'une classification que d'apporter une proximité sémantique entre les éléments du vocabulaire.

On possède la connaissance qu'une « tumeur maligne du colon ascendant » est proche d'une « tumeur maligne du colon descendant » par le seul fait que les codes C18.2 et C18.6 sont frères dans la hiérarchie et donc fils du même père [C18 Tumeur maligne du colon].

Mais une distance sémantique basée sur la seule hiérarchie des codes est très fruste. Reprenant l'exemple précédent, on peut noter qu'il existe le code K57.2, très éloigné des codes C18.- dans la hiérarchie, qui signifie « diverticulose du colon, avec perforation et abcès ». Le mot « colon », commun aux deux libellés, montre qu'il existe un lien très fort entre eux malgré cette distance. La complexité de la connaissance médicale est faite de toutes ces relations.

On retrouve cela dans la notion de concept primaire et de concept composite : « colon » est un concept primaire de type anatomique ; « tumeur du colon » est déjà un concept composite construit à partir de l'anomalie anatomopathologique « tumeur » et de la localisation anatomique « colon ». L'ensemble constitue bien ce que l'on appelle couramment une maladie. La « diverticulose du colon, avec perforation » (sous-entendue du colon) est également une maladie avec un niveau de complexité un peu plus grand que la précédente, la perforation étant une complication non explicitée en tant que telle.

Le fait que la CIM soit traduite en plusieurs langues génère de très nombreux synonymes. C'est un point fort. Lors de la constitution à partir de la CIM-10 du lexique d'une base de connaissance médicale linguistique mise en œuvre pour indexer et interroger du texte, cet aspect a été souligné par l'équipe de Baud [8]. En fait, la conclusion que l'on peut tirer de cette analyse de la CIM valide leur travail, si besoin était : beaucoup de vocabulaire et de nombreuses traductions, c'est un bon point de départ pour construire un thesaurus de termes médicaux. Mais on est encore bien loin de la notion de maladie.

### LES MALADIES DANS LES RUM

Dans les bases PMSI, la notion de maladie pour un patient donné peut *ne pas exister* : un résumé de séjour codé [Z51.0 Séance de radiothérapie] présente un grand silence en terme d'information à son sujet, la radiothérapie étant utilisée aujourd'hui pour d'autres affections que le cancer.

Elle peut être *noyée dans du bruit* : dans un séjour codé [Z090 Examen de contrôle après traitement pour une affection, sans précision], [R53 Malaise et fatigue] et [E119 Diabète sucré non insulino-dépendant, sans complication], seul le dernier code apporte une réelle information de type maladie, les deux premiers pouvant être considérés comme du bruit.

En fait, il est très difficile de savoir ce qui est réellement du bruit et la transition avec le cas suivant n'est pas toujours évidente : *maladie éclatée entre les différents*

*éléments de sa description.* Par exemple, les codes [D69.5 Thrombopénie secondaire], [C78.0 Tumeur maligne secondaire du poumon], [C50.9 Tumeur maligne du sein], [Y42.6 Antigonadotrophines, anti-œstrogènes, anti-androgènes, non classés ailleurs], [R64 Cachexie], coexistent dans un ensemble de résumés pour décrire les vicissitudes de la chimiothérapie dans les cancers du sein évolués, situation fréquente dans nos hôpitaux.

### RETROUVER LA MALADIE

Pouvoir utiliser les bases de résumés médicaux du PMSI pour faire des analyses concernant les maladies serait un progrès important dans le domaine de l'information. La perspective d'un chaînage des résumés par patient incite à se pencher sur la question. D'un point de vue de santé publique, la description doit être suffisamment macroscopique pour présenter un réel intérêt épidémiologique. Dans cet objectif là, il faut résoudre principalement deux problèmes : éliminer le « bruit » lié à l'objectif économique du recueil des données pour que l'essentiel de l'information sur les pathologies puisse être individualisé et trouver le bon niveau de granularité de la description : quels concepts regrouper pour couvrir quelle réalité ?

#### Éliminer le bruit

Cela peut être obtenu, en partie, en ne tenant compte que des codes pouvant bénéficier d'un des types sémantiques définis dans la partie « Vocabulaire et types sémantiques ».

#### Regrouper les éléments de description

Le *tableau III* reprend l'exemple clinique précédent, un résumé de sortie contenant les cinq codes pertinents dans un objectif de description des maladies. Dans ce cas précis, la maladie « centrale » est « le cancer primitif du sein » ; mais le concept qui présente de l'intérêt serait plutôt « le cancer » qui englobe le cancer primitif, sa complication évolutive le cancer secondaire, les complications liées à la thérapeutique... Cela nous ramène au schéma familial à l'étudiant en médecine, légèrement aménagé, présenté dans le *tableau IV*, de la maladie avec sa symptomatologie, ses complications...

On peut partir de l'idée que l'intitulé de la maladie qu'on recherche est un libellé au choix dans la CIM et sa définition l'ensemble des codes qui lui sont attachés structurellement (tous ses fils dans la hiérarchie telle qu'elle existe). Les autres facettes correspondent aussi à des ensembles de codes qui peuvent être rattachés au modèle initial lorsqu'il existe des éléments explicites pour le faire. Ceux-ci sont les inclusions, les exclusions et les caractères

**Tableau III. Exemple d'un ensemble de codes attachés à un patient.**

| Libellés des codes  | Type sémantiques associés                                 |
|---|---|
| [D69.5 Thrombopénie secondaire]   | Maladie simple<br>Complication liée au traitement médical |
| [C78.0 Tumeur maligne secondaire du poumon]   | Maladie simple<br>Complication liée à une maladie         |
| [C50.9 Tumeur maligne du sein]  | Maladie simple  |
| [Y42.6 Antigonadotrophines, anti-œstrogènes, anti-androgènes, non classés ailleurs] | Notion de thérapeutique<br>Indrogène                      |
| [R64 Cachexie]  | Signe ou symptôme   |

**Tableau IV. Concept maladie.**

|                     |   |
|---------------------|---|
| Intitulé            |   |
| Définition          |   |
| Manifestation       | Signe ou symptôme<br>Maladie primitivement dite                                   |
| Complication        | Liée à la maladie<br>Liée au traitement médical<br>Liée au traitement chirurgical |
| Séquelle            |   |
| Étiologie           |   |
| Notion d'antécédent |   |

dagues et astérisques principalement. Mais, par ailleurs, toute une connaissance pourrait être déduite avec des méthodes de type analyse du langage naturel. Enfin, il existe un dernier gisement potentiel de connaissance qui est l'analyse statistique des bases de données qui peut faire émerger des co-occurrences fréquentes de codes qui pourraient être soumises à l'appréciation d'experts.

### RÉSULTATS

La base des RUM de l'année 1998 d'un hôpital universitaire de 900 lits a été complétée par l'adjonction d'un numéro d'identifiant patient connu par ailleurs du système d'information. À titre d'exemple, pour la pathologie cancéreuse, on sélectionne les patients ayant eu au cours de l'année 1998 au moins une fois un code entrant dans leur définition (cf. partie « Regrouper les éléments de description »). On isole l'ensemble des RUM attachés à ces patients. Les codes contenus dans ces RUM vont remplir tout ou partie des facettes de la maladie ou être éliminés. On obtient une maladie, avec ses caractéristiques, par patient.

Pour constituer la maladie « cancer », il faut d'abord savoir que le terme de cancer est synonyme de tumeur maligne, premier obstacle à une automatisation... Le groupe des tumeurs malignes (C00-C97) comprend des

**Tableau V. Maladie cancer.**

| Initialé                               | Tumeurs malignes                                    |
|--|---|
| Définition                             | C00 à C97   |
| Manifestation : signe, symptôme        |   |
| Manifestation : maladie                |   |
| Complication de la maladie             | C00 à C78 et C81 à C97 tumeurs malignes (primaires) |
| Complication du traitement médical     | C77 à C80 tumeurs malignes secondaires              |
| Complication du traitement chirurgical | Y426 Y431 Y432 Y433 chimiothérapies                 |
| Étiologie                              |   |
| Antécédent                             | Z85 antécédents personnels de tumeur maligne        |

sous-groupes correspondant à des tumeurs malignes primitives et secondaires. La présence de l'adjectif « secondaire » dans les libellés permet d'identifier les métastases, terme qui n'est pas connu de la CIM. Les tumeurs primitives sont ainsi définies par exclusion des secondaires parmi les tumeurs malignes. La connaissance qu'une tumeur secondaire est une complication d'une tumeur primitive n'existe pas dans la CIM. De même, le fait que la chimiothérapie est un traitement du cancer. Le *tableau V* illustre la représentation qui peut être faite de la maladie cancer.

Pour 40 de ces patients (sur 3 248 dans la base de l'hôpital), ayant généré 100 RUM correspondant à 98 passages hospitaliers (séjours ou séances), il existe 221 codes CIM. Soixante-cinq d'entre eux (30 %) n'ont pas de rapport avec une pathologie ; 107 (48 %) sont expliqués par la notion de cancer telle qu'elle a été définie. Sur les 49 codes restants, [D70 Agranulocytose], [R64 Cachexie] et [G40.4 Syndromes épileptiques particuliers] coexistent avec un code de cancer et lui sont clairement liés. Cette connaissance ne peut être puisée directement dans la CIM. Elle peut être donnée soit à dire d'expert, soit émerger de l'analyse des bases de données. Les 47 autres codes représentent des comorbidités vraies : [I10 Hypertension artérielle], [I64 Accident vasculaire cérébral] [J45.0 Asthme à prédominance allergique]...

À un niveau très générique, on décrit ainsi 40 patients cancéreux : 35 non-complicés et cinq compliqués de métastases et/ou de problèmes liés à la chimiothérapie. Bien entendu, une description au niveau plus précis, au niveau de l'organe, reste toujours possible (*tableau VI*).

Ce type de description qui extrait, de 221 codes CIM, 40 maladies en apportant une information sur leur stade évolutif et la fréquence des problèmes iatrogènes, approche la notion de maladie dans une optique médicale épidémiologique. Le modèle initial est imparfait : la réalité montre l'existence de pathologies reliées de façon directe

**Tableau VI. Patients cancéreux.**

| Libellé de la maladie : cancer | Effectif non compliqué | Effectif avec complication                                   |
|--------------------------------|------------------------|--|
| ORL                            | 3                      |  |
| Digestif                       | 2                      | Un avec métastase  |
| Respiratoire/thoracique        | 5                      |  |
| Os                             | 1                      |  |
| Tissus mous                    | 1                      |  |
| Peau                           | 1                      |  |
| Appareil génital masculin      | 4                      | Un avec métastase et complications liées à la chimiothérapie |
| Voies urinaires                |                        | Un avec métastase  |
| Système nerveux central        | 7                      | Un avec complication liée à la chimiothérapie                |
| Glandes endocrines             |                        |  |
| Hémopathies malignes           | 6                      | Un avec métastase  |
|                                | 5                      |  |

(épilepsie) ou indirecte (agranulocytose) qui n'avaient pas été prises en compte.

Il faut bien noter à propos de cet exemple que les analyses portent sur un très petit sous-ensemble de la base de l'hôpital : 40 patients cancéreux sur 3 248 de la base. Elles seraient plus démonstratives sur des espaces et des temps importants (hôpital, région, suivi année par année, etc.), mais elle n'ont pas d'autre ambition que d'illustrer le propos concernant la notion de maladie. À noter aussi qu'il s'agit de décrire des maladies et non des malades.

L'objectif de la réflexion proposée n'est pas de définir des liens de causes à effets de type physiologique entre les entités, mais plutôt d'explicitier une occurrence de plusieurs codes tels qu'ils ont été définis par ailleurs (dans la CIM), au sein d'un ensemble de résumés de séjours hospitaliers, de reconnaître et d'isoler ce sous-ensemble parmi tous les codes présents. Comme le font remarquer McCray et Nelson [9] à propos de UMLS où les concepts du metathésaurus se construisent et prennent forme au fur et à mesure de l'intégration de nouvelles terminologies dans le système, on peut imaginer ainsi définir des profils de maladies au sens large par leurs éléments constitutifs codés explicitement dans les bases PMSI, le modèle initial suggéré par la CIM s'étoffant par diverses méthodes au fur et à mesure de son utilisation.

## CONCLUSION

L'analyse de la CIM montre qu'elle n'est quasiment pas utilisable comme mode de représentation de la connaissance des maladies en dehors de son objectif statistique. Par ailleurs, les règles de codages liées à l'objectif médico-économique influent sur l'information sélectionnée : diagnostic renfermant le type de la prise en charge, effet pervers liés aux points Isa... Malgré cela, il est peut être possible d'obtenir assez simplement une idée

des maladies rencontrées à l'aide de « fonctions de recherche » au sein de la base de RSA. Cette modeste réflexion propose des éléments pour y parvenir. Des outils de ce type entraîneraient une homogénéité (donc comparabilité) dans les résultats d'analyses aujourd'hui complètement hétéroclites. Pour l'avenir, cela aurait certainement un effet bénéfique sur la qualité de la description des maladies à proprement parler.

### RÉFÉRENCES

- 1 Ingenerf J, Giere W. Concept-oriented standardization and statistics-oriented classifications: continuing the classification versus nomenclature controversy. *Meth Inform Med* 1998 ; 37 : 527-39.
- 2 Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Meth Inform Med* 1998 ; 37 : 394-403.
- 3 Zweigenbaum P. Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Informations de Santé, Innovation, Stratégie* 1999 ; 2-3 : 27-47.
- 4 Classification statistique internationale des maladies et des problèmes de santé connexes. Dixième révision. Genève : Éditions Organisation mondiale de la santé ; 1995.
- 5 Le PMSI : analyse médico-économique de l'activité hospitalière. Direction des Hôpitaux, éd. *La lettre d'informations hospitalières* ; numéro spécial, mai 1996.
- 6 Botti G, Burgun A, Le Beux P, Fieschi M. Représentation sémantique pour le partage d'information. *Journées émois 99* ; Nancy ; 24-26 mars 1999.
- 7 Surjan G. Questions on validity of international classification of diseases-coded diagnoses. *Int J Med Inform* 1999 ; 54 : 77-95.
- 8 Baud RH, Lovis C, Rassinoux AM, Scherrer JR. Alternative ways for knowledge collection, indexing and robust language retrieval. *Meth Inform Med* 1998 ; 37 : 315-26.
- 9 McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Meth Inform Med* 1995 ; 34 : 193-201.